

THE NEWSPAPER GAME

Automatic Identification of Text Genres to Enhance Language Learning

Debora Ramalho Barros, Carlo Emmanoel T. de Oliveira
Carla Verônica M. Marques and Cláudia Lage Motta
Institute of Information Technology, University Federal of Rio de Janeiro
Av. Brigadeiro Trompowsky, s/n, Rio de Janeiro, Brazil

Keywords: Text genres, Literacy, Naive bayes, Information retrieval, Genre features, Text linguistic, Genre classes, Interdisciplinary.

Abstract: Genre has been receiving a lot of thought in Brazilian schools, due to its great importance on language learning. According to the Brazilian Education Boarder the work with genre and text production is a possible way to solve its issue on high rate functioning illiteracy. This paper aims to discuss ways of implementing an intelligent-web based system that analyses characteristics of 8 textual genres found in the newspaper support media. During the game, the player can build a newspaper and exercise writing different genres, while a Bayesian agent recognizes genres.

1 INTRODUCTION

Genre has been receiving a lot of thought in Brazilian schools (Marcuschi, 2008), due to its great importance on language learning. According to the Brazilian Education Boarder the work with genre and text production is a possible way to solve its issue on high rate functioning illiteracy (Soares, 1998). It's been proved that the more you are in contact with different types of texts either producing or reading them, the more illiterate you become. Hence, we came up with this project of creating a text classifier sort of game with the purpose of helping the production of different genre texts so as to create context and opportunity for freer writing practice as well as providing feedbacks on the writing performance, which shall initially only concern the genre compositional structure.

Firstly, we view text genre or the style of text as characterizing the purpose for which the text has been written. Examples for genre are: research article, novel, poem, news article, editorial, homepage, advertisement, manual, court decision etc. Text-based applications have become more increased, different aspects of text, such as genre, can prove useful for various purposes. Not to mention that, characterizing text differently than the usual subject or prepositional content, has been the focus of many information retrieval and

classification research. In this article we address the issue on automatic detection of the genre class of text.

Genre classes are clearly different from subject classes that most classification research has dealt with. Even though a set of documents may belong to the same class because they share the common topic, they often times serve different purposes, falling into diverse genre classes. As such, classifying documents based on genre would result in a totally different outcome than that from ordinary subject-based classification. From the traditional information Retrieval point of view, a retrieval query about a certain topic such as "Sports" would retrieve many documents related to many different sources of things when submitted to an Internet search engine, but they may be of different genre, such as a sport's TV channel page, sports news, product advertisement, or critical review of a certain game. Genre provides a new dimension for text retrieval and classification, in addition to topicality, and help users become more familiar with the intrinsic structure required in different texts formats.

Automatic genre classification has been studied in the recent past by Bretan *et al.* (1997), Dewe *et al.* (1998), and Dillon *et al.* (2000). Karlgren and Cutting (1994) explored the use of structural cues and rather simple cues such as counts of third person pronouns in text with discriminating analysis. In

subsequent work (1996) she investigated the relationship between the genre of retrieved versus un-retrieved documents and relevant versus non-relevant documents. Used features are simple statistics, such as sentence length and word length, and syntactic complexity such as average depth of a parse tree. Identifying text genre would be beneficial to many text-based applications. For instance, if the genre of every document is known a priori, information retrieval results could be better presented to the user, depending on the preference the user has. As pointed out by Kessler *et al.* (1997), the performance of many natural language processing tools, such as part of- speech tagging, parsing, and word sense disambiguation, could be enhanced since some language usages embedded in grammatical constructions and word senses are related to the genre of text. In Web applications, genre detection would help wrappers that attempt to extract specified information from semi structured. . Kessler *et al.* (1997) identified cues in four categories: structural cues (e.g. counts of POS tags), lexical cues (e.g. words used in expressing dates), character-level cues (e.g. punctuation marks), and derivative cues (e.g. average sentence length as a ration and standard deviation in sentence length as a variation). They decided not to use the structural cues because of the high computational cost. Their computational methods were logistic regression and neural networks (a simple perception and multi-layer perception) that combine 55 cues.

More recently, Stamatatos *et al* (2000) reported on their work for genre detection using word frequencies and punctuation marks. Instead of using sophisticated linguistic cues, they attempted to develop a method that works for unrestricted text in any domain and language with minimal computational cost in extracting cues.

Lee and Myaeng (2004) took the stance more related to traditional information retrieval and text categorization approaches than to deep natural language processing for genre identification. Our text analysis is based on their work.

Features generated from the text analysis are used for genre-based classification of documents. Our approach is similar to the one of Lee and Myaeng (2004) in that they developed their own deviation-based statistical feature selection method utilizing subject-based classification information of the training documents and we are following their steps.

Having developed a genre classifier, they began to investigate on the issue of how text genres help classifying documents based on the subject content

of documents. This is a corollary to our hypothesis that subject classification would help identifying the genre class of a document automatically. Thus helping to enhance writing by providing a feedback according to the genre they choose to write.

Some experimental results are provided, but the work included in this paper is quite exploratory in nature and is still on research.

2 THE GENRE CLASSIFIER

2.1 Overall Method for Feature Extraction

The genre classifier we used is no different from the traditional learning-based classifier. The learner extracts features representing genre classes from training documents whose genre classes are known already, and a classification algorithm determines to which class a new document should belong using the learned representations of the genre classes. In comparison with previous genre classification approaches, furthermore, we agree with Lee and Myaeng (2004) that the difference lies in the types of features used as well as the extraction method itself. The major difference lies in the method by which features are extracted and their weights are calculated, but it is still on research, and we are unable to contribute to it in this paper.

The feature extraction method was derived from text linguistics theory and observation that the frequency of a feature (e.g. a noun, a connector, or verb tense suffixes, etc ...) may be high in a set of document belonging to a genre class, because it represents the particular genre class. This phenomenon is likely to happen especially when the training documents are collected randomly from the entire document space.

As such our feature selection method shall use the statistics from two different class sets, genre classes and subject classes, in the training data. The weight of a feature for a genre will be computed based on three factors Lee and Myaeng (2004):

- ✓ how many training documents belonging to the genre contain the feature (test 1)
- ✓ how evenly the feature is distributed among the subject classes that divide the genre class (test 2),
- ✓ how discriminating the feature is among different genre classes (test 3).

With the first two factors Lee and Myaeng (2004) found features that are found in as many genre documents as possible and distributed as evenly as

possible among all the subject classes that divide the training documents in the genre class. Doing so Lee and Myaeng (2004) have found out that a good genre-revealing feature show up across different subject classes even if a feature appears in many documents belonging to a particular genre class. Furthermore, according to Lee and Myaeng (2004) the third factor ensures good features are as specific to a genre class as possible by downgrading the features that happen to occur in several genre classes. Hence, our experiment counts with the characteristics implied to the genres we decided to work with.

As it is an interdisciplinary research project, we counted on a text Linguistic approach towards the classifier, taking into consideration a reliable source of information taken from recognized text linguistics theories and published data.

In an attempt to teach the classifier what's fundamentally determine each genre type, so that it shall enables a more reliable return of deviating characteristics when a text is tested. Our hypothesis is that if trained documents based on genre class help classifying documents we are able to help students enhance their writing skills by giving them a feedback on their work according to the genre they choose to write.

The feature extraction method was derived from a sieve for seven deliberately chosen pre-established genres, which has been created to calibrate the results given by the Bayes analyses.

2.2 Computation

So far is this research we have done nothing more simply than store as many genre samples. From the period of November to August 2010 we limit the source by using online newspaper as a reference to gather data for analysis. According to Heckerman (1996) Bayesian classification needs a huge amount of data in order to provide as accurate result as possible, bearing in mind the fact that even the largest amount of perfect data entry has only been proven to be 80% solid, Hence what has been measured by us is enough to show how relevant the result is to the purpose of the game. In order to gather different genre text data we used *Crawling and indexing content*, that is the process by which the system accesses and parses content and its properties, sometimes called metadata, to build a content index from which search queries can be served. Meanwhile, we transformed the scrutiny of the selected genres, which in fact, corresponds to a contribution of our research. We organized a matrix

for the dimensions for classifying text genres into taxonomy of genres according to their definitions found in different sources of teaching materials and dictionaries. Moreover, it aids in the construction of attributes for the solidification of the students' texts Bayesian classification.

2.2.1 Features

We created a variable grammatical structure based on one of the six criteria taken from Marcushi (2008), often used to name genres in Brazilian Portuguese language and complemented by our research of genre definitions and concepts taken from various sources. According to the definition of those six criteria we organized the content from the definition of those seven genres into grids, each one corresponding to a dimension, which shall represent a weight after the experiments are concluded.

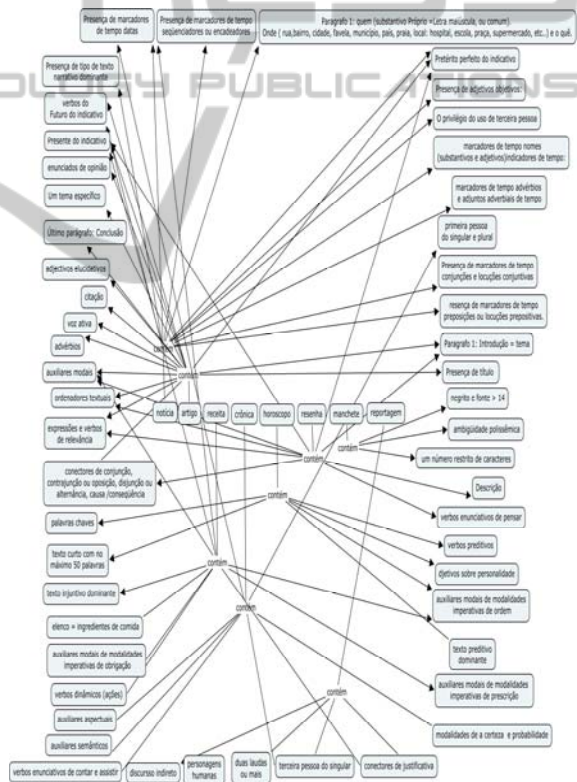


Photo 1: Features of classification.

3 EXPERIMENT

This section reports on our first testing of the initial Hypothesis that if the learner extracts features representing genre classes from training documents

whose genre classes are known already, a classification algorithm will determine to which class a new document should belong using the learned representations of the genre classes.

Unfortunately the feature extraction method is not ready yet we are still running experiments to see to what extent several different types of features (photo1) such as nouns, pronouns, proper nouns, verb endings, exclamations, and special characters are useful for genre-based classification.

The experiment results you shall see below represents the first phase of testing counting only on a simple naïve Bayes approach, which helped us visualize the areas in which the system should be calibrated and extemporized.

3.1 Testing Ground

The documents we used for training and testing were collected from the Web2 in seven genre classes: article, chronicle, critical review, horoscope, news report, recipe and headlines. The collected documents were classified into subject categories using the hierarchy to which they were assigned in the Web sites and the results were examined manually to correct possible errors. The total numbers of Portuguese documents collected are 5,000. The documents collected by a linguist and a computer scientist.

Due to lack of options available just 10 portal sites were used in the collection building process to eliminate possible bias toward document types determined by the Web sites. Each document was examined by two people for inclusion in the collection as well as in the designated genre and subject classes. A half of the collected documents in

Portuguese was used for training and the other half for testing.

Grid 1 shows the numbers of documents in each genre class which have been tested.

Genre classes	number
articles	375
Reviews	141
chronicle	137
horoscope	1395
recipe	1286
headlines	1360
news report	1621

Grid 1: Stored text data.

Effectiveness has been measured by using real text data texts taken form a research done with

students in a Brazilian Language school during the first semester of 2010. As you see in the grid below, not all of the genres were produced, due to varied reasons such as the lack of knowledge of structure and content of some genres, which they have not seen in school yet.

Their texts were applied to the learner that extracted the features and provided results, which determined to which class a new document should belong using the learned representations of the genre classes.

Children's texts	number
articles	0
Reviews	2
chronicle	0
horoscope	3
recipe	3
headlines	10
News report	10

Grid 2: Number of text produced by students.

3.2 Overall Effectiveness

The first experiment was only made to see how effective is the direct application of the Naïve Bayesian approach to genre-based classification. Moreover, it has proved it not at all very reliable, even though it has provided some very accurate results. However, we are aware of the fact that Lee and Myaeng (2004) feature selection method is an alternative of research in order to calibrate these results, and enabling the game to help on the production of different genre texts so as to create context and opportunity for freer writing practice as well as providing feedbacks on the writing performance, which shall initially only concerns the genre compositional structure.

3.3 Results

Surprisingly, just three genres had their results around 50%, as you can see in the grip below. Therefore, this lead us to mistrust the whole experiment itself, due to the fact that having satisfying results to some of the genres may have been just fortune and we should not rely on fortune on scientific experiments. Results, data and figures must be accurate and by no means refuted.

The classification algorithm which determines to which class a new document should belong it must be rather known and calibrated using the learned representations of the genre classes we provide it

with, so as to being able to automatic return a feedback with the features the text either should or shouldn't have present.

Children's texts	number
articles	X
Reviews	0.5444
chronicle	X
horoscope	0.455
recipe	0.5222
headlines	0.999
news report	0.8766

Grid 3: Results.

An experiment which aims at helping text production and complement classroom work, should not only be able to understand the criteria which its results are taken from, in other words, how the naïve bayes algorithm was able to imply what genre class a text belongs. But also explain why it belongs.

All in all this experiment has been the first step of a long research project which shall next apply Lee and Myaeng (2004) method in order to interfere and guarantee a guided result to tests and results by calibrating the results with the features shown in photo 1, which has been the investigative work of a linguist and shall grant this experiment to more accurate results.

4 ROLES OF GENRE IN LANGUAGE LEARNING

The linguistic fundaments in this research is also based on the proposed National Curriculum Parameters (PCNs) to support the teaching of language, both oral and written, in the genres of speech, which has triggered several studies aimed at describing a considerable amount of genres from heterogeneous texts as well as providing suggestions for teaching using texts as examples and reference sources for a particular genre. For example, the book edited by Helen N. Brandão (2000), *Genres of discourse in school* or the various theories about gender and their learning, among which we mention: Deborah C. P. Costa (2001), *The construction of secondary genres in kindergarten: the emergence of genres and news entry*. Campinas: UNICAMP, Daniella L. Days (2001), *Interview by computers: a proposal to analyze the configuration of gender*, Belo Horizonte: PUC; Lusinete V. de Souza (2001),

The achievements of children: the ordinary lines of the text view, São Paulo: PUC. (Kleiman, 2010).

Regarding the Portuguese language teaching the PCNs propose that texts should be worked out according to the axis USE = REFLECTION = USE, aiming to "enable the student to expand the use of language in private bodies in order to use it effectively in public offices knowing how to take on the word and produce texts, both oral and written, coherent, cohesive, appropriate to their recipients, the objectives will be proposed" (1997, p.41). Thus the teaching unit to be considered is the text in its various forms, called by the PCN, gender towards empirical texts (1998, p.13).

The PCN emphasize the responsibility of all disciplines to teach students to use texts that make use a more systematic approach (1997, p.31) However, when we refer to the work with genres, including presenting a remarkable distribution cycles (1997 p.111-112, 128-129, 1998 p.49, 52) The PCNs are not self-explanatory: "A competent writer is someone who, when producing a speech, knowing the possibilities that are offered culturally, know how to select the genre in which his speech will take place by choosing whichever is appropriate to their objectives and circumstances described in question. For example, if what you want is to convince the reader, the writer responsible selects a genre that allows it to produce a predominantly argumentative text, whether it is making a request to a particular authority, probably draft a letter (...)" (1997, p.65). That is, the document contained only the statement that the teacher shall choose the most appropriate genre, since the Curriculum is not a textbook, it works rather like a guideline, helping teachers towards the understanding of the light concepts of linguistic theories as to the didactic in the classroom in order to develop activities using texts in the classroom.

The *Newspaper game* was designed as an attempt to allow the execution of a task in education in general, and in the classroom in particular, that can address the gender perspective in the discussion here and lead students to generate and analyze the most diverse linguistic events, whether written or oral, and identify the characteristics of each genre. Although always suggested by the literature and rarely applied in the classroom, due to its high demand of time, effort and cost, the game works as an exercise which, besides instructive, also allows the text production practice.

We believe that production will enable the student to create a daily newspaper that implicitly do reflect some of the genre, such as its main

characteristics in terms of content, composition, style, language level and purposes. Clearly, this task can be rephrased in many ways, according to the intersets of each teaching situation. However, we believe that it is more modest for the analysis; it will always be very promising. And even more promising, will enable the student to the practice of producing these genres through the game's newspaper.

Working with genre is an extraordinary opportunity to deal with language in its various daily uses. "Within a linguistic perspective everything we do will be done at some genre. So everything we do can be treated linguistically in either gender" (Marcuschi, 1996). The game of the newspaper brings seven genres that appear in various media that, are produced systematically and with great impact on daily life, without excluding the virtual media to be worked, the internet and computer. An analysis of textbooks of language teaching materials shows that there are a variety of texts types present in these works. However, when we look more closely reveals that the variety does not match an analytic reality.

However, we believe that students learn naturally by producing the various genres written in everyday use. As it is common to naturally learn the more formal oral genres, as well observe Joaquim Dolz and Bernard Schneuwly (1998). Thus, there is not an ideal type of genre for teaching language. But it is likely that one can identify progressive difficulties with gender, level of less formal to more formal, more private to more public and so on. We decided to choose the genres at random, just observing the universe of genres found in the printed newspaper in our country today.

4.1 The Game

The game platform was developed according to the web platform called *Phidias* (Carla et al, 2009a) and, therefore, designed with three environments: two interfaces one for the student and the other to the applicator (teacher), to monitor and receive evaluations of the process through Bayesian algorithms, which contains measures of cognitive requirements (screens) to help the diagnosis and mapping of the process of textual development. At the child's interface, a newspaper similar to the printed version, composed with empty text boxes corresponding to the seven genres studied here, along with the specification of the text required at each box. The genres were carefully selected from a variety of genres found in real newspapers here in

Brazil. Each text produced shall be subjected to a Bayesian analysis which must return a percentage result of how faithful to the genre is its text, resulting in a diagnosis on the state of knowledge you have each genre linguistic composition and structure.



Photo 2: The game layout of the child's interface.

The intervention of the applicant (teacher) occurs after the diagnosed from the results the bayes analysis, through a method known as *Fio Condutor*; in which the teacher evaluates and rehabilitates mentally the child through continue use of the *Newspaper game*. During the sessions, the child can go through seven stages as shown in photo 3 below.

FASES DE CONSTRUÇÃO E APLICAÇÃO DO PROTOCOLO/JOGO						
VERSÃO 1	VERSÃO 2	VERSÃO 3	VERSÃO 4	VERSÃO 5	VERSÃO 6	VERSÃO 7
AVALIÇÃO ESTÁTICA	AVALIÇÃO DINÂMICA	AVALIÇÃO DINÂMICA	AVALIÇÃO DINÂMICA	AVALIÇÃO ESTÁTICA	INTERVENÇÃO	REPETIÇÃO DAS VERSÕES 1 E 2
APRESENTAÇÃO DO ESTÍMULO	MEDIAÇÃO PARA A CONSTRUÇÃO DA REGRA	CONSTRUÇÃO DA REGRA COM OBJETOS MANIPULÁVEIS	ELABORAÇÃO DIRIGIDA	RETESTE MEDIADO		COM TAREFAS DIFERENTES

Photo 3: Fio Condutor (Marques, 2009).

In the first phase, the teacher introduces the game and the student can explore the possibilities in part 1, the proposal is that student tells what they see. In the second everything on screen can move and they build their own newspaper, while the applicant (teacher) observes and asks children to report what they did. This is a first attempt to the student reflection about their activity. In a third instance, the child can move across the screen and rebuilt it and fell free to write the texts. In version 4, called 'Elaboração dirigida', created by psychologist Seminério (1997), the user is subjected to a discursive and reflective questionnaire based on the results provided. In step 5, a new game with a new

set of genres. This project will serve until the fifth stage, and the subsequent remain as future proposals.

4.2 Theoretical Framework

The conceptual contribution lies in two main topics related to this study: assessment instruments and cognitive development, and textual typology. At first, we point out the theory by Franco Lo Presti Metaprocessual Seminário (1997), developed based on the following theories. A. Bandura (1980) Modeling (Transmission Model or rules) as a means of promoting learning, which is a result of stocking about models that do not need to be strengthened at the time of purchase, Bruner (1966) generative power or value in order to generate new hypotheses and combinations. Also called generative rules. Noam Chomsky (1968) the rule of recursion is the innate basis for the development of logic and recursion. Flavell (1963) the metacognitive strategy and comprehensive model of cognitive control, Gestalt (1945) dynamism inherent in the cognitive structuring the state of cognition to metacognition through insight. Vygotsky (1930) participation in teacher learning and social environment.

5 CONCLUSIONS

In this paper we show the failure of an Bayesian analysis experiment which has only helped us to invest in a better methodology for genre classification using taxonomy and statistics rather than just the naïve Bayes approach. The proposed method uses concepts from selected genre-revealing features from the textual Linguistic literature. The deviation formula will make use of both genre-classified and conceptual features to eliminate features that can interfere in the classification and return useful information, so that teachers will be able to intervene in their students learning processes in a more effective way.

All in all we plan to have this mechanism running by the end of this year so that we undergo school experiments to validate the game next year.

ACKNOWLEDGEMENTS

Our thanks go to Rackel Reis and Maíta Carvalho, Allan Valente and Yago for either helping us with the programming part of the experiment as well as Lee and Myaeng for orienting us with information

retrieve and the Natural language process state of art available to complement our research and methods thus introducing us to a great method.

REFERENCES

- Moore, R., Lopes, J., 1999. Paper templates. In *TEMPLATE'06, 1st International Conference on Template Production*. SciTePress.
- Smith, J., 1998. *The book*, The publishing company. London, 2nd edition.
- Wolkert, Jussi Karlgren, "Web-Specific Genre. Visualization", Proc. of the 30th Hawaii International Conference on System Science, Jan 1997.
- Johan Dewe, Jussi Karlgren, Ivan Bretan, "Assembling a Balanced Corpus from the Internet", 11th Nordic Conference of Computational Linguistics, pages 100-107, Copenhagen, 1998.
- Andrew Dillon, Barbara A. Gushrowski, "Genre and the Web: Is the Personal Home Page the First Uniquely Digital Genre?", *JASIS*, 51(2):202-205, 2000.
- Jussi Karlgren, "Stylistic Variation in an Information Retrieval Experiment", Proc. of the 2nd International Conference on New Methods in Language Processing-NeMLaP, 1996.
- Jussi Karlgren, Ivan Bretan, Johan Dewe, Anders Hallberg, Niklas Wolkert, "Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres", 8th DELOS Workshop on User Interfaces in Digital Libraries, pages 85-92, 1998.
- Jussi Karlgren, Douglass Cutting, "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis", Proc. of COLING94, Kyoto, 1994.
- Brett Kessler, Geoffrey Nunberg, Hinrich Schotze, "Automatic Detection of Text Genre", ACL'97, pages 32-38, July 1997.
- Yong-Bae Lee and Sung Hyon Myaeng, "Text Genre Classification with Genre-Revealing and Subject-Revealing Features," Proceedings of the 25th ACM SIGIR Conference, pages 145-150, Tampere, Finland.
- D. Lewis and M. Ringuette, "Comparison of two learning algorithms for text categorization," Proc. Of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.
- H. J. Oh, S. H. Myaeng, and M. Lee, "A practical hypertext categorization method using links and incrementally available class information", Proc. of the 23rd ACM SIGIR Conference, pages 264-271, Athens, Greece, 2000.
- E. Stamatatos, N. Fakotakis, G. Kokkinakis, "Text Genre Detection Using Common Word Frequencies", Proc. of the 18th International Conference on COLING2000, 2000.
- Y. Yang and X. Liu, "A re-examination of text categorization methods," Proc. Of the 22nd ACM SIGIR Conference, 1999.
- Marcushi, Luiz Antônio, "Produção textual, análise de gênero e compreensão." Parábola, 2008.

- Marques, C. V. M. (2009). *Laboratório de neuropsicologia cognitiva-projeto geral: avaliação de crianças deficientes visuais*. Rio de Janeiro: NCE/UFRJ, 10 p. (Relatório Técnico, 02/09).
- Marques, C. V. M; Motta, C. L. R; Oliveira, C. E. T.; Vrabl, S. D. P.; Lapolli, F., Pereira, A. P. M; Daflon, L. (2009a). *Avaliação de Crianças Deficientes Visuais através de jogos neuropedagógicos. SCA 2009 - Simpósio de Computação Aplicada*, Passo Fundo, Rio Grande do Sul.
- Marques, C. V. ; Oliveira, C. E. T. ; Motta, C. (Org.). *A revolução cognitiva; um estudo sobre a teoria de Franco Lo Presti* Seminário. Rio de Janeiro: PPGI/IM/NCE, (2009b). (Relatório Técnico, 04/09).
- Travaglia, L. C. (2007) *A caracterização de categorias de textos: tipos, gêneros e espécies*. *Alfa: Revista de Linguística*, v. 51, p. 39-79, 2007. ISSN/ISBN: 19815794.
- PCN - *Parâmetros curriculares nacionais: língua portuguesa* (1998). Brasília: MEC/SEF.
- Seminário, F. L. P. (1997). *Novos rumos na Psicologia e na Pedagogia: metacognição: uma nova opção*. *Arquivos Brasileiros de Psicologia*, Rio de Janeiro, v. 49, n. 3, p. 5-22, jul./set.
- Soares, Magda. *Letramento: um tema em três gêneros*. 2ed. Belo Horizonte: Autêntica, 2000.